Identification and protection of Personal Data using blockchain technology

Roman Dorosh[1]

1. Department of Information Technology, Vasyl Stefanyk Precapathian National University, Ukraine,
Ivano-Frankivsk, 57 Shevchenko Street,
E-mail: roman.dorosh.17@pnu.edu.ua

Анотація - Блокчейн діятиме як прозора технологія проти несанкціонованого доступу через систему соціальних і смарт-контрактів. Вищі навчальні заклади можуть використовувати технологію блокчейн для різноманітної академічної діяльності, наприклад для адміністрування онлайн-курсів. Майже кожна сфера викладання та навчання має потенціал отримати вигоду від технології Blockchain. Також блокчейн чудово вирішує проблему шахрайства в цілому завдяки своїм здібностям – децентралізації, узгодженості та здатності діяти як глобальна система.

Ключові слова – ідентифікація, захист, блокчейн, смарт-контракти, освіта, NLP, Ethereum.

Introduction

Blockchain provides the main characteristics: autonomy, decentralization, fault-tolerance, distribution, open, immutability and transparency [1], which led to interest of researchers into this area;

This paper proposes the high-level overview of the system, the plagiarism detection algorithm with multilanguage support, the smart contract which deployed into test EVM (Ethereum Virtual Machine). Also, this paper shows real-world outcomes, in-detail description of the comparing algorithm using blockchain and SimHash, comparing to other approaches and future vision of this initial approach.

Related Work

So, the common approach to capture the plagiarism is to apply NLP and found similarity score between selected documents. The Anzelmi et al [1] goes this way. There is lots of papers which using text mining to find the information and algorithms like cosine similarity [2] to compare them. Researchers also used Machine Learning algorithms with combination of text mining approach [4]. Machine Learning and particularly Neural Networks can be applied to such areas as software defined demodulation of weak radio signals, which proposed in [5] papers and efficient identity document detection [7] etc. Research [8] shows the importance of introducing new ways to improve education, especially online engineering education.

In this paper, the SimHash [10] algorithm is used, because of its performance based on billion pages.

Implementation

SimHash Approach

The main advantage of this algorithm is scalability and speed. The original paper [8] shows practically usefulness for identifying near-duplicates in documents belonging to large file storage (multi-billion repository).

So, based on implementation areas and attributes of SimHash, this algorithm was chosen as engine for plagiarism detection. In this approach, we look through set of documents S = {D1, D2, D3, … Dn} and find SimHash of every entry SHd = SimHash(D). After that, every entry hash SHd compares to all documents S one by one by corresponding distance between them. The result of this operation will be number between [0, 1] which can be translated to corresponding percentage. SimHash algorithm gives power to cluster similar documents into groups and detect plagiarism of many entries by one query to the database. The main disadvantage here that algorithm cannot show the accurate places where the cheater copies from the original paper.

Text Processing and NLP

The first step is text extracting from *PDF, DOC, DOCX, PPTX* formats. In this paper, online text mining is currently not used, so research only focused on manually adding test documents to test Ethereum blockchain.

The preprocessing algorithm performs the following steps: Tokenizer – Removing stop words – Stemming – Lemmatization – Conversion to word embeddings.

Ethereum Blockchain Smart Contract

In this paper, smart contract is not complicated because it’s only need to store simhashes of the documents and sha256 hash on blockchain. Solidity is used and Ganache as a local blockchain. I propose the smart contract written in Solidity with two functions: add SimHash and sha256 hash of document and retrieve all documents.

New Unified Algorithm

In this paper I propose the unification of existing plagiarism detection methods, i.e. simhash and decentralized blockchain technology into one algorithm (Figure 1). The addition of cross-language plagiarism detection also proposed, but not fully implemented yet.

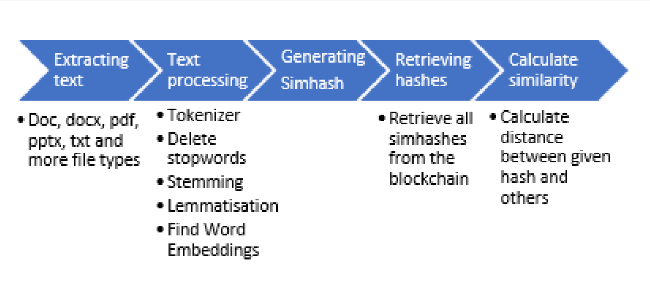
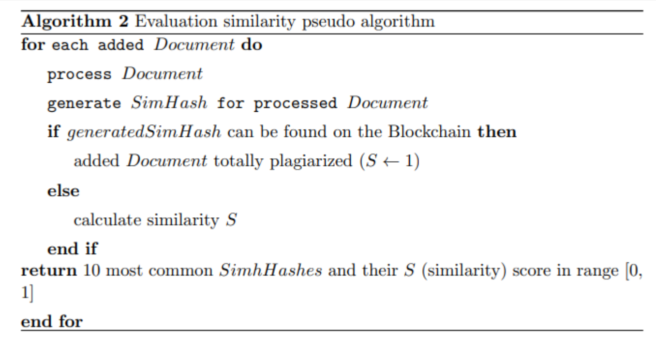


Figure 1 – New Unified Algorithm

Evaluating similarity with unified algorithm

In this paper, I have implemented the evaluation similarity pseudo algorithm:



```
Algorithm 2 Evaluation similarity pseudo algorithm
for each added Document do
    process Document
    generate SimHash for processed Document
    if generatedSimHash can be found on the Blockchain then
        added Document totally plagiarized ($S \leftarrow 1$)
    else
        calculate similarity $S$
    end if
return 10 most common SimhHashes and their $S$ (similarity) score in range [0, 1]
end for
```

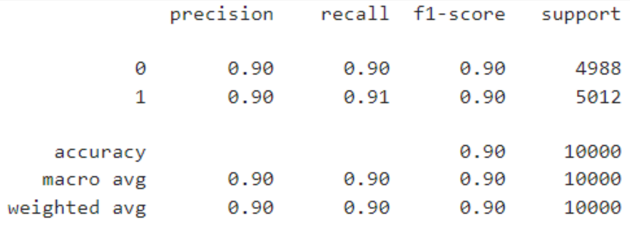
Figure 2 – Pseudo algorithm

Evaluation of experimental results

The dataset [10] for evaluation proposed in this paper consists of 143,000 news articles from 15 major publications, but used only 10,000. I can formalize this task, as a classification with two classes (plagiarism - 0 and not plagiarism - 1).

I have performed testing on Processor 11th Gen Intel(R) Core(TM) i7-1185G7 @ 3.00GHz, 2995 Mhz, 4 Core(s), 8 Logical Processor(s) with 32 GB RAM and 512 GB SSD drive, OS is Windows 10. The Python library for visualizing metric results *sklearn* was used.

So, the results of testing my algorithm is given below on the Figure 3 and Figure 4



```
              precision    recall  f1-score   support

           0       0.90      0.90      0.90      4988
           1       0.90      0.91      0.90      5012

    accuracy                           0.90     10000
   macro avg       0.90      0.90      0.90     10000
weighted avg       0.90      0.90      0.90     10000
```

Figure 3 – New algorithm metrics

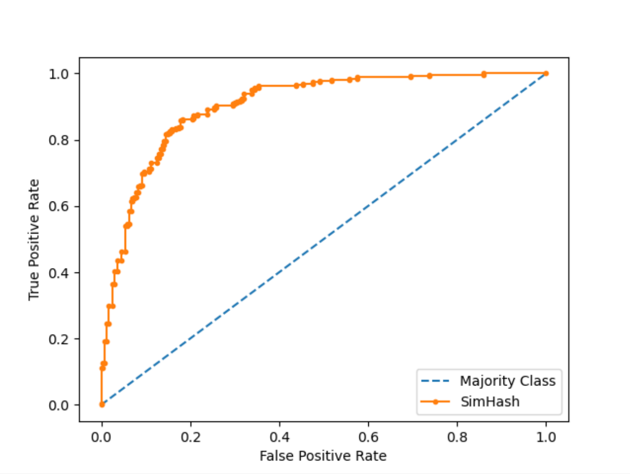


Figure 4 – ROC AUC score,

After comparing new algorithm to TF-IDF algorithm the overall performance was better by 7%.

Conclusion

The proposed unified algorithm operates well on detecting exact or partial copies of the documents; also the performance doesn't downgrade with scalability and size of the documents; changing words to their synonyms or paraphrasing can be detected too; identification of the multi-language plagiarism performs on the basic level, to achieve more precise solution, the complex language analysis must be conducted.

References

[1] M. S. Ali, M. Vecchio, M. Pincheira, K. Dolui, F. Antonelli and M. H. Rehmani, "Applications of Blockchains in the Internet of Things: A Comprehensive Survey," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1676-1717, Secondquarter 2019, doi: 10.1109/COMST.2018.2886932.

[2] Anzelmi, Daniele, Domenico Carlone, Fabio Rizzello, Robert Thomsen and Dil Muhammad Akbar Hussain. “Plagiarism Detection Based on SCAM Algorithm.” *IMECS 2011* (2011).

[3] Anu Saini, Ankita Bahl, Supriya Kumari and Mitali Singh. Article: Plagiarism Checker: Text Mining. *International Journal of Computer Applications* 134(3):8-11, January 2016. Published by Foundation of Computer Science (FCS), NY, USA.

[4] Chitra, A. and Rajkumar, Anupriya. "Plagiarism Detection Using Machine Learning-Based Paraphrase Recognizer" *Journal of Intelligent Systems*, vol. 25, no. 3, 2016, pp. 351-359. https://doi.org/10.1515/jisys-2014-0146.

[5] M. Kozlenko, I. Lazarovych, V. Tkachuk and V. Vialkova, "Software Demodulation of Weak Radio Signals using Convolutional Neural Network," *2020 IEEE 7th International Conference on Energy Smart Systems (ESS)*, 2020, pp. 339-342, doi: 10.1109/ESS50319.2020.9160035.

[6] M. Kozlenko and V. Vialkova, "Software Defined Demodulation of Multiple Frequency Shift Keying with Dense Neural Network for Weak Signal Communications," *2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET)*, 2020, pp. 590-595, doi: 10.1109/TCSET49122.2020.235501.

[7] M. Kozlenko,V. Sendetskyi, O. Simkiv, N. Savchenko, A. Bosyi, “Identity Documents Recognition and Detection using Semantic Segmentation with Convolutional Neural Network (short paper)”. *Cybersecurity Providing in Information and Telecommunication Systems 2021 (CPITS)*, 2021, pp. 234-242.

[8] Dutchak, M., Kozlenko, M., Lazarovych, I., Lazarovych, N., Pikuliak, M., Savka, I. “Methods and Software Tools for Automated Synthesis of Adaptive Learning Trajectory in Intelligent Online Learning Management Systems”. *Innovations in Smart Cities Applications Volume 4. SCA 2020. Lecture Notes in Networks and Systems*, vol 183. Springer, Cham. https://doi.org/10.1007/978-3-030-66840-2_16.

[9] G.S. Manku, A. Jain, A.D. Sarma, “Detecting Near-Duplicates for Web Crawling”, research.google.com, https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33026.pdf (accessed Aug. 23, 2022).

[10] “All the news dataset”. Kaggle.com https://www.kaggle.com/datasets/snapcrack/all-the-news (accessed Aug. 27, 2022)